Comparative Sensor Fusion Between Hyperspectral and Multispectral Satellite Sensors for Monitoring Microcystin Distribution in Lake Erie

Ni-Bin Chang, Benjamin Vannah, and Y. Jeffrey Yang

***Abstract*—Urban growth and agricultural production have caused an influx of nutrients into Lake Erie, leading to eutrophication in the water body. These conditions result in the formation of algal blooms, some of which are toxic due to the presence of Microcystis (a cyanobacteria), which produces the hepatotoxin microcystin. The hepatotoxin microcystin threatens human health and the ecosystem, and it is a concern for water treatment plants using the lake water as a tap water source. This study demonstrates the prototype of a near real-time early warning system using integrated data fusion and mining (IDFM) techniques with the aid of both hyperspectral (MERIS) and multispectral (MODIS and Landsat) satellite sensors to determine spatiotemporal microcystin concentrations in Lake Erie. In the proposed IDFM, the MODIS images with high temporal resolution are fused with the MERIS and Landsat images with higher spatial resolution to create synthetic images on a daily basis. The spatiotemporal distributions of microcystin within western Lake Erie were then reconstructed using the band data from the fused products with machine learning or data mining techniques such as genetic programming (GP) models. The performance of the data mining models derived using fused hyperspectral and fused multispectral sensor data are quantified using four statistical indices. These data mining models were further compared with traditional two-band models in terms of microcystin prediction accuracy. This study confirmed that GP models outperformed traditional two-band models, and additional spectral reflectance data offered by hyperspectral sensors produces a noticeable increase in the prediction accuracy especially in the range of low microcystin concentrations.**

***Index Terms*—Harmful algal bloom, image fusion, machine learning, microcystin, remote sensing.**

I. INTRODUCTION

HUMAN POPULATION growth and agricultural land use have inevitably led to an increase in eutrophic conditions in surface waters. The subsequent influx of nutrients has fueled cyanobacteria-dominated algal blooms in polluted waters in many parts of the globe [1]. Blooms containing toxins that negatively impact human health and the environment are referred to as harmful algal blooms (HABs). Not only can HABs form and spread rapidly but wind and water currents will mobilize the blooms [2]. The dynamic movement of the HABs requires constant monitoring and forecasting, due to the threat posed to humans recreating on the lake, commercial fishing operations, and water treatment facilities. The predominant species of cyanobacteria that produce cyanotoxins are *Microcystis aeruginsa*, *Microcystis viridis*, *Aphanizomenon flos-aquqe*, and *Anabaena*. While there are a variety of cyanotoxins, microcystin is the main toxin produced [1], [3]. The aberrant toxicity of microcystin can lead to liver cancer, liver failure, and even death [1], [4]. To ensure the protection of human health from microcystin exposure, it is necessary to develop a reliable method for the near real-time prediction of microcystin within hazardous algal blooms.

Satellites can provide medium to high-resolution images of selected light spectra. Using the detected surface reflectance emissions, predictions of microcystin concentrations become possible. The theoretical basis for this claim lies behind the fact that every substance gives off a unique spectral signature. As a substance is exposed to different portions of the electromagnetic spectrum, it will reflect a certain percentage of the light. The percentage of reflectance can be plotted as a function of wavelength to clearly display which frequencies have an affinity for absorbing and reflecting. The unique curve produced is known as a spectral signature. The substance would have the defining spectral peaks and troughs, almost like a fingerprint for that object, where much of the radiation has either been reflected or absorbed. The intensity of the reflectance at different wavelengths can then be used to determine the amount of the substance present in the water. However, the relationship between reflectance and concentration is highly nonlinear for certain substances. As a result, effective data mining techniques must be applied to accurately predict the concentration for an observed spectral response. In this paper, we demonstrate the utility, technical difficulties, as well as data mining approaches for near real-time monitoring of microcystin concentrations in Lake Erie.

II. LITERATURE REVIEW

The prediction of microcystin concentrations in a lake poses a unique problem, since 95% of the microcystin is contained within healthy *Microcystis* cells [5]. It is not until death or induced rupture of the cell wall that the toxin is released. Thus, in order to generate

Manuscript received October 03, 2013; revised April 20, 2014; accepted June 01, 2014. Date of publication July 08, 2014; date of current version August 01, 2014. This work is a part of U.S. EPA Water Resources Adaptation Program (WRAP) supported by EPA contract under Grant EP-C-11-006. This paper has been subjected to the administrative review by EPA and has been approved for external publication. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Agency; therefore, no official endorsement should be inferred. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

N.-B. Chang and B. Vannah are with the Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: nchang@ucf.edu; benvannah@gmail.com).

Y. J. Yang is with the U.S. EPA, Office of Research and Development, National Risk Management Research Laboratory, Cincinnati, OH 45268 USA (e-mail: Yang.Jeff@epamail.epa.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2014.2329913

an accurate estimate of microcystin concentration, it is necessary to establish a relationship between microcystin and other substances present in the water. These substances may serve as indicators of microcystin concentration. Chlorophyll-a levels in *Microcystis* blooms are related to the amount of microcystin present [1], [6], [7]. Since *Microcystis* is a bacterium that uses photosynthesis for energy production, it is reasonable to conclude that high concentrations of *Microcystis* can be linked with elevated chlorophyll-a levels. In a study by Budd *et al.* [8], Advanced Very High-Resolution Radiometer (AVHRR) and Landsat Thematic Mapper (TM) images were used to determine chlorophyll-a concentrations in the lake, leading to detect and track algal blooms. Their study established that it is possible to use surface reflectance data to detect and track algal blooms based upon chlorophyll-a levels, and Wynne *et al.* [9] expanded the depth of this type of study by using the surface reflectance of chlorophyll-a to specifically predict *Microcystis* blooms, instead of algal blooms in general. It was discovered that *Microcystis* blooms can be distinguished from other cyanobacteria blooms through close analysis of the detected surface reflectance at 681 nm [10]. Studies by Mole *et al.* [11] and Ha *et al.* [12] had similar findings in regard to using chlorophyll-a as an indicator for quantifying microcystin in algae blooms that have stabilized and reached the late exponential growth phase and stationary phase. In summary, chlorophyll-a is a reliable indicator of microcystin for *Microcystis* HABs that are no longer in the peak of the exponential growth phase.

Phycocyanin is a pigment that all cyanobacteria contain [1], and it has been shown that phycocyanin concentrations share a positive correlation with microcystin levels [7]. In a study by Vincent *et al.* [13], Landsat TM images in the visible and infrared spectral bands were used to generate algorithms to predict phycocyanin concentrations with prediction accuracies from 73.8% to 77.4%. Thus, the surface reflectance of phycocyanin, chlorophyll-a, and *Microcystis* are suitable indicators for the prediction of microcystin levels in a lake. The surface reflectance curves for chlorophyll-a and phycocyanin in surface waters peak at 525, 625, 680, and 720 nm. Landsat and MODIS (multispectral fusion pair) jointly capture half of the peak at 625 nm and half of the peak at 680 nm, while MERIS and MODIS (hyperspectral fusion pair) jointly detect the full spectral feature at 680 nm. The multispectral fusion pair also has bands in the shortwave infrared range (band centers at 1650 and 2090 nm), which shed insight into chlorophyll concentrations in the water. As the concentration of chlorophyll increases, the absorbance decreases [14]. The hyperspectral fusion pair is capable of detecting electromagnetic radiation centered at 412 and 443 nm; 412 nm is associated with the degradation products of once living organisms [15]. This is important for toxic microcystin prediction, because the toxin is only released from the *Microcystis* bacteria once the organism dies. Finally, 443 nm is associated with chlorophyll maximum absorption [15].

While Landsat and MERIS can detect spectral features for chlorophyll-a and phycocyanin, a significant drawback is their 16 and 3 day revisit cycles. The daily revisit time of the MODIS sensor can fill in the data gaps through the use of data fusion. However, MODIS alone cannot be used as a substitute because of its poor spatial resolution (250/500 m) for the land bands, which is outclassed by Landsat with 30 m spatial resolution, and its 1000 m spatial resolution for the ocean bands, which is surpassed by MERIS with 300 m spatial resolution. We propose an ultimate solution by fusing Landsat and MODIS (MODIS' land bands) or MERIS and MODIS (MODIS' ocean color bands) pairwise to generate a synthetic image with both enhanced spatial and temporal resolutions. Such a synthetic image can enable near real-time monitoring of microcystin concentrations, thereby populating a database with information on spatial occurrences, areas prone to HAB formulation, general movement patterns of HABs on the lake, and seasonal maps. This information provides water treatment, fishing operations, and areal residents with the knowledge required in decision-making.

Having presented the motivation for fusing the selected satellites images, the next consideration is given to the intercomparisons between multispectral and hyperspectral sensors. Multispectral sensors collect a smaller number of noncontiguous, wide spectral bands (less than 20) [15]–[19]; they typically offer enhanced spatial resolution. Hyperspectral sensors, on the other hand, provide a greater spectral resolution by capturing numerous spectral bands with a bandwidth of 10–20 nm. Since their emergence in the 1970 s and 1980 s, hyperspectral remote sensing techniques have advanced significantly with the development of the Compact Airborne Spectrographic Imager (CASI) in 1978 and the proposal of the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) to the National Aeronautics and Space Administration (NASA) in 1983. Due to the small width of each band, hyperspectral sensors have a worse signal-to-noise ratio than multispectral scanners; the latter collects more photons per band, which lessens the impact of the noise [20].

Multispectral sensors aptly function in the open ocean, because the surface reflectance is the combination of a few water quality parameters at the water's surface [21]. Case 2 water bodies (such as examined in this study) exhibit significantly more optical complexity than the open ocean, and the algorithms utilizing multispectral data sensor products demonstrate reduced performance in these waters [22], [23]. The benefit of hyperspectral sensors is the number of additional bands that they provide at a thinner bandwidth. The added bands with more narrow bandwidths more accurately depict the spectral reflectance curve of the water body, as is presented in Fig. 1.

From Fig. 1, one can see that the defining peaks and troughs for Landsat are smoothed out when the total reflectance is averaged for the bands; therefore, the resulting band is unable to reveal detailed reflectance information that may be necessary for discriminating between water quality constituents and characterizing the species within a phytoplankton bloom. In comparison, MERIS captures the unique spectral features (peaks and troughs). The application of multispectral and hyperspectral sensor imagery for ocean color remote sensing has been reported by Lubac *et al.* [25], which concluded that surface reflectance data from multispectral and hyperspectral sensors can be used to quantify phytoplankton blooms, yet an enhanced hyperspectral resolution provides superior quantitative assessment and monitoring of phytoplankton blooms. Through this approach, the superior detail of hyperspectral information introduces more degrees of freedom, and allows for optical models and algorithms of higher explanatory power to quantify the nonlinear relationships between surface reflectance and concentrations, more accurately classifying species and concentrations of water quality

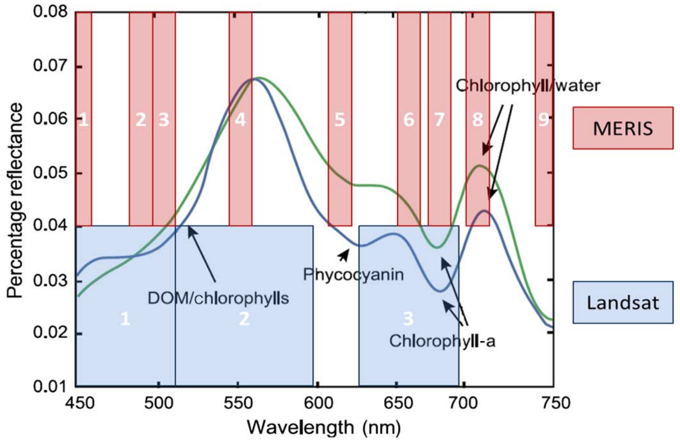

Fig. 1. Comparing the bandwidths between multispectral (Landsat) bands and hyperspectral (MERIS) bands (adapted from [13]).

constituents, and enhancing the determination of inherent optical properties that vary with depth [20], [26].

The goal of this study was to develop the integrated data fusion and mining (IDFM) technique of combing surface reflectance data from multispectral and hyperspectral sensors, and further to quantify their performance for providing near real-time monitoring of the spatiotemporal distributions of microcystin in a lake. In addition to real-time monitoring, seasonal maps of microcystin were retrieved to assess HAB spatial distributions throughout the year. In this paper, we wish to explore: 1) the feasibility in predicting microcystin concentrations in a lake using the IDFM technique; 2) which of the fused band combinations are most useful in determining microcystin concentrations in a case 2 inland water body; and 3) whether hyperspectral sensor products provide a significant advantage over products from multispectral sensors for microcystin prediction.

III. Methodology

A. Study Site

Lake Erie is one of the five Great Lakes located in North America. Together, the lakes make up the largest supply of fresh water in the world. The lakes provide drinking water for over 40 million Americans, in addition to 56 billion gallons per day withdrawn from the lakes for industrial, agricultural, and municipal use [2]. Each summer, the Great Lakes are threatened by *Microcystis* blooms, yet the blooms in western Lake Erie are the most severe and contain levels of microcystin that are not suitable for drinking water. Throughout the 2000 s, *Microcystis* blooms have increased in frequency and severity [24], [27].

B. Satellites Used and In Situ Data

Surface reflectance data utilized in this study were obtained from Landsat TM, MERIS, and MODIS sensors. MERIS is a hyperspectral sensor with a moderate 300 m resolution (reduced resolution products are available at 1.2 km spatial resolution) and has a 3-day revisit time for sites near the equator. While the MERIS sensor itself has hyperspectral capabilities, the processing onboard the satellite yield a product set that is considered multispectral. This is because the MERIS sensor acquires hyperspectral data from 390 to 1040 nm at a 1.8 nm spectral resolution for a total of 520 bands, and unwanted spectral data is disposed of, while the remaining spectral values are averaged into 15 discrete bands [15]. The band widths and locations are programmable, and the 15 product bands of MERIS still offer greater detail about the visible, near-infrared, and infrared frequencies than a number of other satellite sensor products available in this range. Landsat offers superior spatial resolution at 30 m; however, only 7 spectral bands are available compared to the 15 of MERIS. The revisit time of Landsat is significantly longer at 16 days. In developing the near real-time monitoring system, daily satellite images of the area of interest are required for data fusion techniques that are used to fill in the data gaps of MERIS and Landsat by using the MODIS sensor which has a daily revisit time. The spatial, temporal, and spectral resolutions of the satellites central to this technical approach are compared in Table I. Data fusion at the pixel level using the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) requires input images to be spectrally similar [28]. In accordance with the requirement, as is shown in Table I, MERIS is fused with the spectrally analogous ocean color bands of MODIS, while Landsat TM bands were fused with the land bands of MODIS.

The National Oceanic and Atmospheric Administration (NOAA) is a provider of the *in situ* data for microcystin concentration. NOAA collects surface water samples in western Lake Erie, when probable blooms are identified based on their analysis of satellite images. Samples are taken at the surface to provide surface microcystin concentrations that coincide with the surface reflectance observed in satellite data products. Total microcystin concentration was quantified using ELISA (Abraxis; 520011) after cell lysis (Abraxis; Quik-lyse kit 529911QL) of unfiltered samples [29]. In total, 44 microcystin measurements, unobstructed from view by cloud cover, were made from 2009 to 2011 and suitable for ground-truth usage (Table II). These data only include those with sampling locations free of cloud-cover, land aerosol contamination, and significant suspended sediment levels in the corresponding satellite images.

C. Methodology

The IDFM technique for the prediction of microcystin is shown in Fig. 2. It is designed to fuse satellite data streams and apply machine learning algorithms to derive a working model relating the data streams to the desired output parameter. For this study, Landsat, MERIS, and MODIS surface reflectance imagery serve as the data streams, and the estimated concentration of the toxin microcystin is the desired output parameter. Data fusion techniques are applied to incorporate data into a single image for analysis by machine learning models, which create prediction models for near real-time monitoring and data gap-filling applications.

The IDFM technique consists of five main steps. Step 1) is the acquisition of the surface reflectance data from MERIS, Landsat, and MODIS. Step 2) formats the images for fusion followed by the application of data fusion techniques and algorithms. This study applied the STARFM algorithm to fuse the MODIS (ocean bands) and MERIS pair, and also the MODIS (land bands) and Landsat pair. In Step 4), the synthetic images and ground-truth data were used as inputs for data mining. A genetic programming (GP) model was trained in Discipulus to create an explicit,

TABLE I
Spatial, Temporal, and Spectral Properties of the Satellite Sensors Used in This Study

Parameters	Hyperspectral sensor pair		Multispectral sensor pair	
	MERIS	Modis Terra (ocean bands)	Landsat TM	Modis Terra (land bands)
Product	MER_FR_2P	MODOCL2	LT5	MODO9
Spatial resolution	300 m	1000 m	30 m	250/500 m
Temporal resolution	1–3 days	1 day	16 days	1 day
Band number: band center ± band width (nm)	1: 412 ± 10	8: 413 ± 15	1: 485 ± 35	3: 469 ± 10
	2: 443 ± 10	9: 443 ± 10	2: 560 ± 40	4: 555 ± 10
	3: 490 ± 10	10: 488 ± 10	3: 660 ± 30	1: 645 ± 25
	4: 510 ± 10		4: 840 ± 60	2: 859 ± 18
		11: 531 ± 10	5: 1650 ± 100	6: 1640 ± 12
	5: 560 ± 10	12: 551 ± 10	7: 2090 ± 130	7: 2130 ± 25
	6: 620 ± 10			
	7: 665 ± 10	13: 667 ± 10		
	8: 681 ± 10	14: 678 ± 10		
	9: 708 ± 10			
	10: 753 ± 10	15: 748 ± 10		
	11: 760 ± 10			
	12: 779 ± 10			
	13: 865 ± 10	16: 869 ± 15		

The band centers shared between the satellites have been aligned in the table. Band combinations that occur on the same row are suitable candidates for spectral fusion.

TABLE II
Ground-Truth Samples Were Taken at Various Sites in Western Lake Erie on These Days

	Jun.	Jul.	Aug.	Sep.
2009		7, 14		
2010	28	26	2, 16, 30	2
2011		12	11	14

nonlinear equation relating the fused band data to the ground-truth data. Finally, Step 5) uses the GP model created in Step 4) to compute microcystin concentration maps using the fused band data generated in Step 3) (Fig. 2).

1) Data Acquisition (Fig. 2; Step 1): Surface reflectance data for Lake Erie were collected from the ENVISAT MERIS, Terra MODIS satellite, and Landsat TM sensors. Level 2, ocean-band images for 2009–2011 from the Terra MODIS satellite were downloaded from the online repository through the NASA Ocean Color Web. Since multiple ground-truth samples were taken at different locations on the same day, the MODIS images were inspected for cloud cover for all locations. The level 2 image was downloaded as an HDF-EOS image, only when at least one location was not obstructed by cloud cover. The same criterion was applied to the rest of the satellite data acquired. Additionally, Level 2, land-band images for Terra MODIS were downloaded from the online repository overseen by the NASA Land Processes Distributed Active Archive Center (LP DAAC), United States Geological Survey (USGS), and Earth Resources Observation and Science Center (EROS). Level 2, full-resolution MERIS data were obtained through the European Space Agency (ESA). Failure of the MERIS sensor in April 2012 prevented the usage of ground-truth data succeeding this event. Finally, the Landsat TM data were obtained through the USGS by way of the Global Visualization Viewer, which is maintained by LP DAAC, USGS, and EROS.

2) Image Processing [Fig. 2; Step 2]: The MODIS images were preprocessed at a level 2 basis. This includes the radiometrically calibrated data that were atmospherically corrected for aerosols and scattering [30]. Full-resolution MERIS data came processed on a level 2 basis, with radiometric, geometric, and atmospheric corrections [15]. Landsat data came processed on a level-1T basis with radiometric and geometric corrections [31]. Because the fusion process requires input image pairs to have the same bit-depth and spatial resolution, the input images were processed in ArcGIS, a mapping and spatial analysis software.

Specifically, MERIS images were processed by the following procedures.

1) Reprojection to the Universal Transverse Mercator (UTM) zone 17 North.
2) Crop out land data surrounding Lake Erie.

The MODIS ocean-band and land-band images were processed similarly.

1) Reprojection to UTM zone 17 North.
2) Ocean bands were resampled to the 300 m spatial resolution to match those of MERIS, and the land-band images were resampled to the 30 m spatial resolution of Landsat.
3) Crop out land data surrounding Lake Erie.
4) Ocean-band surface reflectance values were converted to integer values using the same offset and scaling that the ESA applies to MERIS data.

Landsat images were processed as follows.

1) Atmospherically correct Landsat images using the LEDAPS processing software available through NASA.
2) Reprojection to UTM zone 17 North.
3) Crop out land data surrounding Lake Erie.

The image processing consists of three categories of actions: 1) modification of the geometric projections, pixel size, bit depth, and scale in order to fuse them properly; 2) atmospheric correction; and 3) preparation of the image to increase the accuracy of the fused image by removing land contamination from the original images. With regard to the first category, the images need to have the same geographic projection and scaling prior to fusion. Otherwise data for the same pixels between the

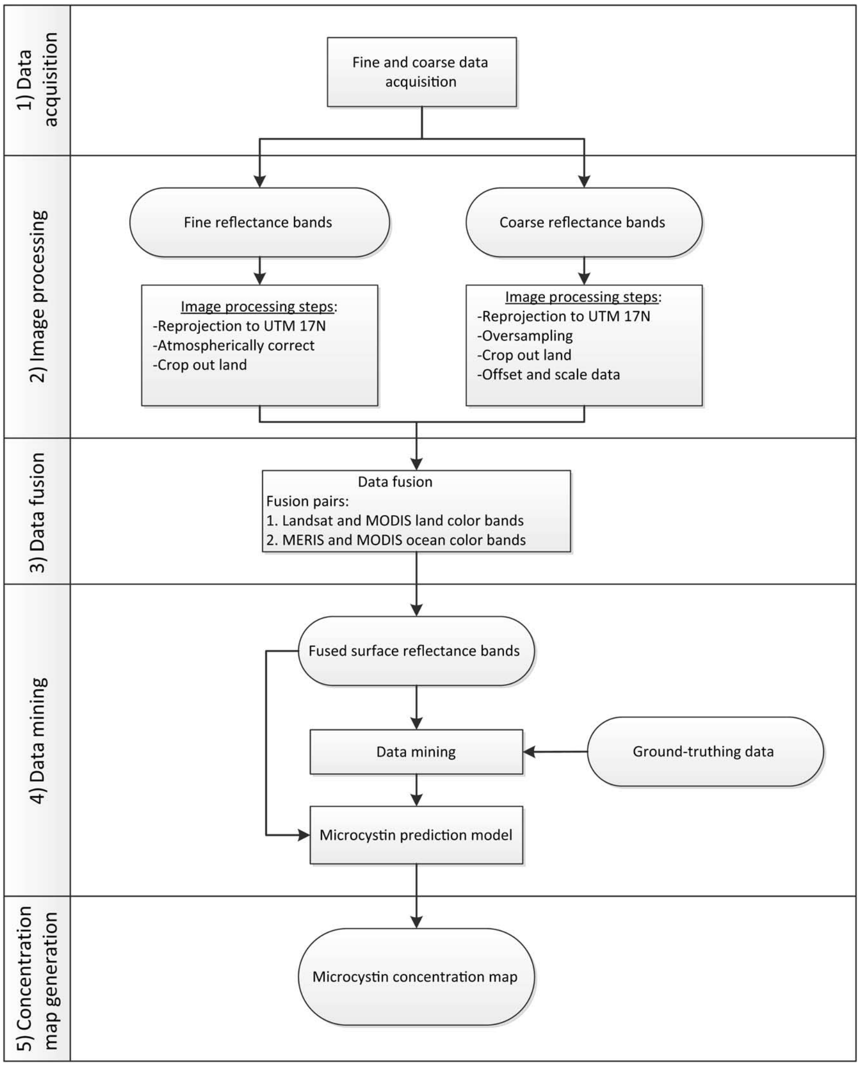

Fig. 2. Methodological flowchart for the IDFM procedure using surface reflectance data from hyperspectral or multispectral sensors.

satellite pairs becomes incomparable, because they represent different swaths of land and have differently scaled values. Additionally, the MODIS ocean-band data were resampled to match the resolution of MERIS and Landsat. The MODIS ocean color surface reflectance values were scaled upward to match those of MERIS, because integer values are required for input into STARFM. This ensures that both input images have the same number of pixels, enabling the pixel-by-pixel comparison techniques used by STARFM. Initially, the Landsat and MODIS data are not the same bit depth, but atmospheric correction using the LEDAPS processing tool scales the Landsat values from 0-255 to -100 to 16 000 [32], [33]. This is because the same MODIS 6S radiative transfer techniques are applied to correct the Landsat data.

For the other two categories, atmospheric correction is needed to remove the scattering effects of the atmosphere from the raw data, thus, producing surface reflectance instead of top of atmosphere radiance. The last category of processing was performed on all images with the purpose to mask the pixel values for the land surrounding Lake Erie. In order to clip the land from all images in ArcGIS, a shape file of Western Lake Erie was used to distinguish between land and water. This step is required to prevent fusing land pixel values with surface water values during processing with the STARFM algorithm.

3) Data Fusion (Stage 3): A fused image is created by the algorithmic fusion of the spectral, temporal, and spatial properties of two or more images [34] (Fig. 2). The resulting synthetic or fused image has all the characteristics of the input images, and incorporates object's defining attributes into a single image with a potential to increase the reliability of the data [35]. Fusion of spatial and temporal properties on a pixel-by-pixel basis was based on the STARFM algorithm by NASA. For this study, the algorithm was used to fill in data gaps caused by the 1- to 3-day revisit time of MERIS using MODIS ocean color bands, and the 16-day revisit time of Landsat using MODIS land color bands. The Landsat and MERIS images are of higher quality than MODIS, but they are sparse in time. As a result, MODIS data are used to capture temporal changes during the periods of data gaps. The overall workflow of the STARFM algorithm is detailed in Fig. 3.

The methodology described here is the fusion of Landsat and MODIS images as the input pair; the same approach is applied for the MERIS (higher spatial resolution image) and MODIS pair. In order to generate a synthetic Landsat image L_0

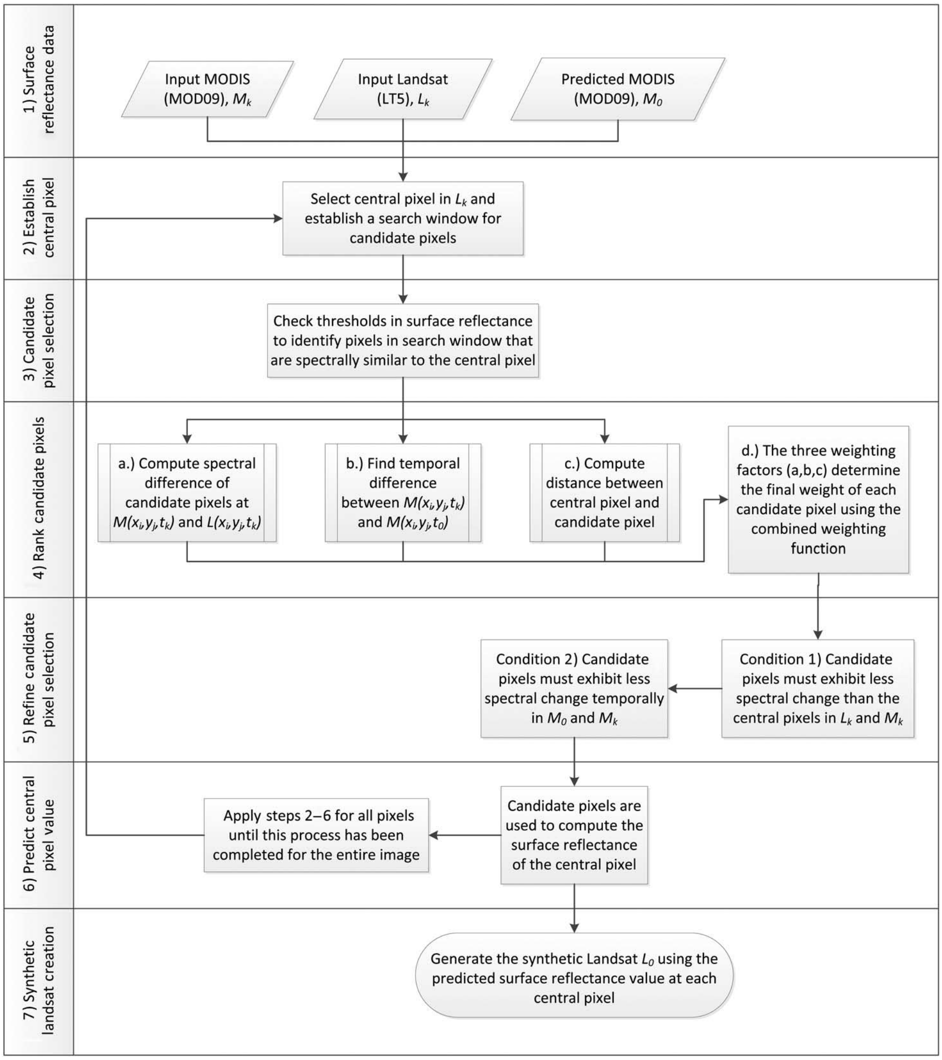
1) Surface reflectance data
Input MODIS (MOD09), M_k
Input Landsat (LT5), L_k
Predicted MODIS (MOD09), M_0
2) Establish central pixel
Select central pixel in L_k and establish a search window for candidate pixels
3) Candidate pixel selection
Check thresholds in surface reflectance to identify pixels in search window that are spectrally similar to the central pixel
4) Rank candidate pixels
a.) Compute spectral difference of candidate pixels at $M(x_i,y_j,t_k)$ and $L(x_i,y_j,t_k)$
b.) Find temporal difference between $M(x_i,y_j,t_k)$ and $M(x_i,y_j,t_0)$
c.) Compute distance between central pixel and candidate pixel
d.) The three weighting factors (a,b,c) determine the final weight of each candidate pixel using the combined weighting function
5) Refine candidate pixel selection
Condition 1) Candidate pixels must exhibit less spectral change than the central pixels in L_k and M_k
Condition 2) Candidate pixels must exhibit less spectral change temporally in M_0 and M_k
6) Predict central pixel value
Candidate pixels are used to compute the surface reflectance of the central pixel
Apply steps 2–6 for all pixels until this process has been completed for the entire image
7) Synthetic landsat creation
Generate the synthetic Landsat L_0 using the predicted surface reflectance value at each central pixel

Fig. 3. Procedural flow for the STARFM algorithm as shown for the MODIS and Landsat fusion pair. The same process is applied when using MERIS (substitute for Landsat) and MODIS ocean color products [28].

(the higher spatial resolution image) to fill in the data gap at time t_0, a MODIS image M_0 from that day is required. This MODIS image is referred to as the predicted MODIS in following discussions. Next, a temporally analogous MODIS M_k and Landsat L_k image pair are required from before or after the prediction date t_0. The input image pair (M_k and L_k) obtained for date t_k serve as a boundary condition detailing how the area of interest looked prior to or after t_0. The dates (t_0 and t_k) of the predicted and input images should be as close as possible, because the chance for significant spectral change between the two images increases with time. In the case of the Landsat input image, it would preferably be from the next 16 day revisit cycle. In this study, the maximum permissible time difference between an input Landsat image and the date for the predicted image was 2 Landsat revisit cycles or less than 32 days. Acquiring these images is Step 1).

Step 2) involves selecting a central pixel from the L_k image. This is a sequential process, which starts with the first pixel in the image and then moves to the second pixel, etc. During each round of iteration, the central pixel value in L_k, M_k, and M_0 is the same. Step 3) identifies the spectrally similar candidate pixels which will be used to predict the reflectance of the central pixel in L_0. Locating pixels that share spectral similarity with the central pixel can be performed using two methods: 1) unsupervised classification; and 2) checking thresholds in surface reflectance [28]. The latter method is used in STARFM. Surface reflectance from the red and near-infrared band is used for this process. The central pixel is designated as the middle class and the standard deviation of the fine resolution pixels in the search box as well as the number of classes are used to categorize the candidate pixels. Pixels sharing similar classification types as the central pixel are then selected as candidate pixels. This method is very similar to traditional classification; however, a classification map cannot be derived for the entire image using the second method. This is because the local rules used to classify pixels within the search window change based upon the new range of pixel values observed as the window is moved around the next central pixel. The user defines the distance of the search box and the number of classes. If the study area consists of heavily mixed, heterogeneous pixels, it may be necessary to increase the size of the search window allow for more homogeneous MODIS pixels to be located. Reducing the number of classes generates lax requirements for candidate pixel selection. The default STARFM parameter values were found to generate reliable results for the study area. A search distance of 750 m was used, and 40 classes were selected for identifying spectrally similar pixels.

Step 4) is designed to rank the candidate pixels based on three criteria that determined how much they are related to the central pixel. From Step 4a), the spectral differences between the candidate pixels in the input MODIS and Landsat are computed, as shown [28]

$$S_{ijk} = \left|L\left(x_i, y_j, t_k\right) - M\left(x_i, y_j, t_k\right)\right| \tag{1}$$

where x_i corresponds to a row in the image, y_j denotes a specific column, $L(x_i, y_j, t_k)$ refers to a specific pixel in the input Landsat image, and $\mathrm{M}(x_i, y_j, t_k)$ refers to a specific pixel in the input MODIS image. In (1), it is assumed that both images have been accurately georeferenced to ensure the supersampled MODIS pixels correctly align with the Landsat pixels. One additional assumption is that if a MODIS pixel $(M(x_i, y_j, t_k))$ and Landsat pixel $(L(x_i, y_j, t_k))$ value for the same day is equal, then the MODIS pixel value $(M(x_i, y_j, t_0))$ from the prediction day equals the synthetic Landsat pixel value $(L(x_i, y_j, t_0))$. If the fine Landsat pixel is spectrally similar to the supersampled MODIS pixel detailing the same area, then $Sijk$ will yield a small value and a higher weighting will be assigned to this pixel location. Obviously, a Landsat and MODIS image taken on the same day should nearly be the same, but this is not always the case, due to slight lags in the overpass timing and differences in spatial resolution. Recall that the spatial resolution of MODIS can be up to 500 m, compared to the 30 m of Landsat. As a result, the spectra of objects within the 500 m are averaged for the MODIS pixel. This is visually depicted in Fig. 4.

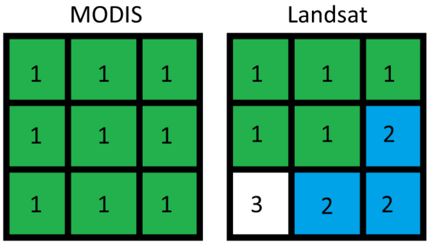

Fig. 4. This is a visual representation of how Landsat is capable of differentiating between objects larger than 30 m in size, while the MODIS image from the same day and for the same location averages the spectral values of all objects smaller than 500 m into a single pixel, due to its lower spatial resolution. The 1 s and green cells denote and algal bloom, the 2 s and blue cells denote open water, and the 3 s and white cells correspond to clouds.

In Fig. 4, the nine Landsat pixels clearly delineate the outline of the bloom from clear water. The MODIS image on the left has been oversampled to match Landsat's 30 m resolution, yet, due to averaging, the observed spectra are most closely related to the algal bloom that covered over half of the pixels in the Landsat image. The purpose of (1) or step 4a) is to determine how well the spectral reflectance of a MODIS pixel compares to the Landsat pixel value at the same location. For example, the green pixels in the upper left corners of the boxes in Fig. 4 are both green and contain the same value. As a result, there is no spectral difference. According to (1), if the two have minimal spectral differences, a small value for S_{ijk} will be computed and a high weighting will be assigned to that candidate pixel. Alternatively, the pixel in the lower right of the MODIS image contains a green square, while the Landsat pixel at this location is colored blue. As a result, the pixels are not spectrally related, and a low weighting would be assigned to the candidate pixel for this location. These examples explain how differences in spatial resolution are resolved by linking pixels at the same location for Landsat and MODIS images taken on the same day. Finally, since the acquisition times for the two satellites are not the same, there is the possibility for clouds to move and cover portions of the image. Consider the pixel in the lower left corner of the images. During the acquisition of the Landsat image, a cloud contaminated this location, but the cloud may not have been present, or it may have been too small to significantly influence the spectral value of the MODIS image. Regardless, (1) computes a low weighting for this candidate pixel, since there is no spectral similarity between MODIS and Landsat pixels.

Step 4b) compares the spectral changes that occur temporally in the input and predicted MODIS images, as detailed [28]

$$T_{ijk} = |M(x_i, y_i, t_k) - M(x_i, y_i, t_0)| \tag{2}$$

where $\mathrm{M}(x_i, y_j, t_0)$ refers to a specific pixel in the predicted MODIS image. A large value of T_{ijk} indicates that there has been significant change in the water quality at this candidate pixel, and it is assigned a lower weighting. The designers of the STARFM algorithm are acting on the assumption that change over time occurs slowly; thus, a drastic change in spectral value at a given location is unlikely. While spectral similarities are initially rewarded, the logistic formula shown in (4) makes the cumulative weighting factor less sensitive to spectral changes [28]. Step 4c) follows the basic logic that candidate pixels closer to the central pixel should receive a higher weighting, as shown [28]

$$D_{ijk} = 1.0 + \frac{\sqrt{\left(x_{w/2} - x_i\right)^2 + \left(y_{w/2} - y_i\right)^2}}{A} \tag{3}$$

where w is the user defined search box side length, $x_{w/2}$ is the row of the central pixel, $y_{w/2}$ is the column of the central pixel, and A is a constant relating the importance of the spatial distance D_{ijk} weight to the spectral S_{ijk} and temporal T_{ijk} weighting factors. As the value for A decreases, the importance of the weight D_{ijk} increases. Candidate and central pixels that are close together will likely exhibit similar spectral changes over time, whereas a candidate pixel farther from the central pixel is less spatially similar and it receives a lower weighting. Step 4d) combines individual ranking criteria (D_{ijk}, S_{ijk}, and T_{ijk}) to form an overall weighting factor for each candidate pixel. This is accomplished in the below equations: [28]

$$C_{ijk} = \ln\left(S_{ijk} * B + 1\right) * \ln\left(T_{ijk} * B + 1\right) * D_{ijk} \tag{4}$$

$$W_{ijk} = \frac{\left(\frac{1}{C_{ijk}}\right)}{\sum_{i=1}^{w} \sum_{i=1}^{w} \sum_{k=1}^{n} \left(\frac{1}{C_{ijk}}\right)} \tag{5}$$

where B is a scale factor and W_{ijk} is the combined weighting factor. The value for B is 10 000 when using LEDAPS reflectance products, and a value of 54 645 was used for the MODIS

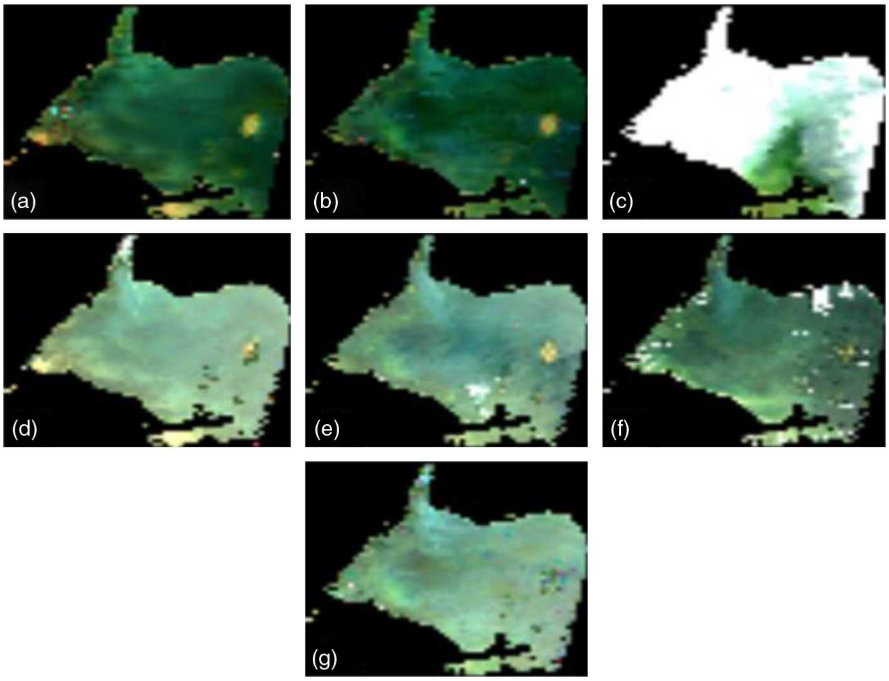

Fig. 5. Gap filling using STARFM. Images (a)–(c) correspond to true color MODIS images, while (d)–(f) are true color Landsat images. Image (g) is the synthetic Landsat generated by STARFM to fill in the hypothetical data gap if image (e) were not available. Images (a) and (d) were taken on the same day; similarly, images (c) and (f) were taken on the same day.

ocean color and MERIS pair, since a scaling factor of 1.83×10^5 is applied to MERIS products to store them as an integer.

Step 5) further refines the selection of candidate pixels based on two conditions shown in the below equations: [28]

$$S_{ijk} < \left|L\left(x_{w/2}, y_{w/2}, t_k\right) - M\left(x_{w/2}, y_{w/2}, t_k\right)\right| \tag{6}$$

$$T_{ijk} < \left|M\left(x_{w/2}, y_{w/2}, t_k\right) - M\left(x_{w/2}, y_{w/2}, t_k\right)\right|. \tag{7}$$

Equation (6) requires that candidate pixels in the input image pair exhibit less spectral change than the central pixels, and (7) requires that candidate pixels in the input and predicted MODIS images show less temporal change than the central pixels. Otherwise, the pixel is considered a "worse neighboring pixel," and it is not used for the predicting the surface reflectance of the central pixel in the synthetic image. Now that a suitable subset of candidate pixels have been related to the central pixel, the predicted surface reflectance for the central pixel in the synthetic image is performed, as shown in Step 6). Steps 2–6) are repeated for all of the pixels. The entire process of summed up in (8) [28]

$$L(t_0) = \sum_{i=1}^{w}\sum_{j=1}^{w}\sum_{k=1}^{n} W_{ijk} * \left[L\left(x_i, y_j, t_k\right) - M\left(x_i, y_j, t_k\right) + M\left(x_i, y_j, t_0\right)\right] \tag{8}$$

where $L(t_0)$ is the synthetic Landsat image formed using spatial information from the fine spatial resolution Landsat image and the temporal changes from the coarse MODIS images. It should be noted that pixels containing clouds cannot be used for the fusion process, and they must be masked out. Furthermore, if there is a great deal of change between the predicted date and a boundary image or one of the boundary images exhibits a significant temporal difference, then the fused results will be less accurate. Finally, this explanation only showed one pair of input images being used to generate the synthetic image. If an additional input pair is used the results can be improved [28]. Think of this as providing the algorithm with a set of both pre and post conditions, instead of just one. This study provided the algorithm with both pre- and postcondition, as long as they were cloud-free and taken within two revisit cycles of the prediction date.

An illustration of the STARFM data fusion procedure for gap filling is provided in Fig. 5. The top row of Fig. 5 is composed of MODIS images, the bottom row is Landsat images. In the event that Fig. 5(e) was not available, then STARFM could be applied to generate a synthetic Landsat to predict spectral reflectance values for that day at 30 m resolution. To do this, a set of preconditions are required. The MODIS in Fig. 5(a) and Landsat in Fig. 5(d) were acquired on the same day, and they provide insight into how the lake looked prior to the data gap [Fig. 5(e)]. The accuracy of the predicted or synthetic Landsat image can be improved by including a set of post condition images. This is provided by MODIS in Fig. 5(c) and Landsat in Fig. 5(f) taken on the same day. The last essential component for gap filling with STARFM is the MODIS in Fig. 5(b). This image provides coarse resolution data for the spectral conditions on the prediction date. Using the three MODIS images in Fig. 5(a)–(c) along with the two Landsat images in Fig. 5(d) and (f), the synthetic Landsat in Fig. 5(g) was generated. Ideally, the true Landsat [Fig. 5(e)] and the predicted Landsat [Fig. 5(g)] would have the same spectral values. The accuracy of the prediction was assessed by computing the coefficient of determination between the RGB images (e) and (f). The coefficient of determination was 0.8278. The true and predicted images are quite similar, but it is important to note that little contribution was made by the image pair in Fig. 5(c) and (f). It is imperative that the input images are free of clouds; however, if clouds are present over a lake, STARFM will not

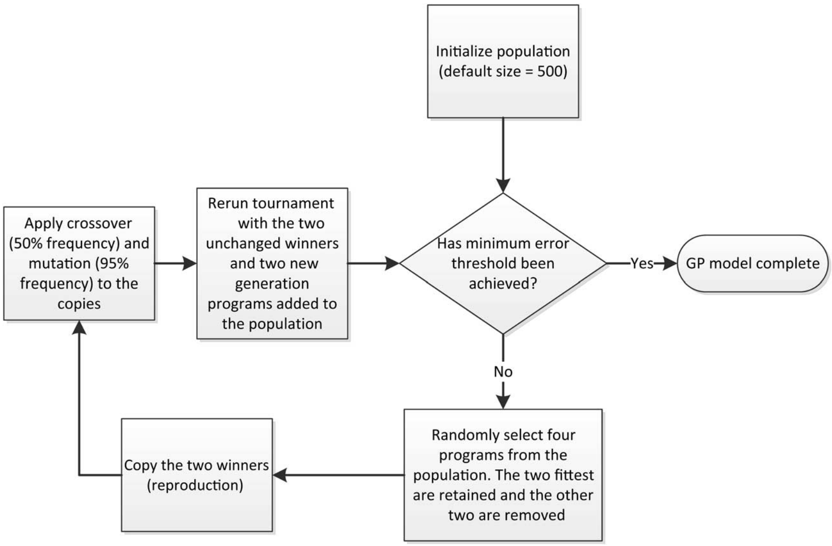
Initialize population (default size = 500)
Has minimum error threshold been achieved?
Yes
GP model complete
No
Randomly select four programs from the population. The two fittest are retained and the other two are removed
Copy the two winners (reproduction)
Apply crossover (50% frequency) and mutation (95% frequency) to the copies
Rerun tournament with the two unchanged winners and two new generation programs added to the population

Fig. 6. Discipulus methodology for training a GP model [36].

mistakenly incorporate cloudy pixels, due to the significant spectral difference from the water pixels.

4) Data Mining [Fig. 2; Step 4]: The IDFM technique permits the use of numerous machine learning approaches to derive an explicit equation or black box model relating the fused surface reflectance data to the ground-truth observations. Notable data mining and machine learning algorithms include GP, Artificial Neural Networks (ANN), ANN and Adaptive Resonance Theory, Constrained Optimization Techniques, Adaptive Dynamic K-means, Principal Component Analysis, and Support Vector Machines. For this study, a GP model was used to relate surface reflectance to microcystin concentration. GP models attempt to establish a relationship between inputs and outputs by applying mathematical, data transfer operations, stack rotation, and conditionals. To begin, GP creates a population of random programs. The larger the population size, the more accurate the resulting model, yet the greater the population size, the longer the training times, and computational requirements [36]. After the initial population is created, programs are ranked by their fitness level. If a program yields an error that is within the minimum error threshold set by the user, then the modeling process is completed. If not, the next step is to use the current suite of programs to develop the next generation. Three primary genetic operators are applied to the current programs to replace poorly performing programs with the new generation of programs.

1) Reproduction: Highly ranked programs are copied without change.
2) Crossover/recombination: Nodes are swapped within a program (asexual recombination), between two programs (sexual recombination), or between more than two programs (multirecombination) to develop the new program.
3) Mutation: Changes are randomly made to the programs with the best fit. This promotes genetic diversity within the population of high ranking programs.

The GP model for this study was created using the Discipulus software package. The user provides the software with inputs and outputs, which are used to train and calibrate the model. During training, the accuracy of a model is determined using least-squares. Discipulus identifies 30 of the best programs, and the model exhibiting the highest fitness is usually selected [36]. The Discipulus methodology is presented in Fig. 6:

Prior to modeling the relationship between certain input and output data using GP techniques, one may wish to ascertain if GP can successfully be used to uncover and model the relationship between datasets in a timely manner and with high accuracy. Predicting the time it takes for GP to develop a reliable solution is not directly possible. However, it is known that the size, integrity, and relationship complexity of the supplied data for analysis are all variables that factor into the time it takes for the GP training process to converge to a solution; and software selection parameters such as program size and mutation rate can further lengthen training times [36]. Predicting solution times is a problematic aspect directly resulting from GP methodology, which features innately random subroutines. First, Discipulus' GP methodology starts by randomly generating programs. Sometimes a few of the randomly generated programs are quite similar to the actual modeling solution required, which would reduce the time necessary to converge to the final model of highest fitness. Additionally, the section of a program that is selected for manipulation by crossover and mutation is random. As a result, it may take numerous iterations until a favorable node of a program is selected for usage with a primary genetic operator or until an unfavorable portion that does not aid in determining the relationship between the input and output data is removed from the program. Fortunately, steps can be taken to reduce training times and expedite the development of a model with high fitness. While one advantage of the GP methodology is to uncover and learn the relation between the input and output data set without any prior knowledge of the problem, one way to save time and increase the rate at which a reliable model is developed is to withhold extraneous input data during the learning process. Thus, the researcher is aiding the model by denying it access to a useless portion of the data, but this is only possible because there is foreknowledge that a portion of the dataset has little impact on predicting the desired output. If provided with the full, unmodified dataset, the GP model should have eventually concluded that certain groups from the input data are extraneous, yet it would take computational effort and time for the GP training process to arrive at this deduction.

The GP models were compared against a traditional two-band model, which was solved through a linear regression model using band ratios instead of individual bands as explanatory variables [13]. The generic setup for a two-band model is shown below

$$C_{MS} = a * \frac{Rrs(\lambda_1)}{Rrs(\lambda_2)} + b \tag{9}$$

where C_{MS} is the concentration of microcystin (signified as the MS subscript) in units of microgram per liter ($\mu\mathrm{gL}^{-1}$), $\mathrm{Rrs}(\lambda)$ is the atmospherically corrected surface reflectance at the band center wavelength, λ is the wavelength of the band center given in units of nanometer (nm). The coefficients a and b denote the slope and intercept obtained through regression. Additionally, a spectral slope two-band model was included in the analysis [37]. The spectral slope is calculated using

$$Slope = \frac{R_{rs}(\lambda_1) - R_{rs}(\lambda_2)}{|\lambda_1 - \lambda_2|}. \tag{10}$$

A nonlinear exponential fit was used to determine the spectral slope coefficients relating the exponential increase in absorption with wavelength for chlorophyll and phycocyanin. For both the models, band combinations were compared to determine the two bands possessing high correlation with chlorophyll-a and phycocyanin estimation (both indicators of *Microcystis*). The same training and calibration data sets used for creating the GP models were employed to train and calibrate the two-band model.

5) Concentration Map Generation [Fig. 2; Step 5]: Microcystin concentration maps for western Lake Erie are generated by applying the GP model to the fused data product created in Step 3). For each pixel of the fused image, there are eight surface reflectance values, one corresponding to each band shared, for MODIS and MERIS, and there are six for MODIS and Landsat. For this study, these band values are used as variables in the explicit equation created from the GP model. As determined by the GP model, certain band values will share a strong relationship in the determination of the microcystin concentration, while others may offer weak explanatory power. Thus, the GP model uses the fused surface reflectance values of the pixel to predict the microcystin concentration at that location. After this process is applied to the entire lake, a clear depiction of microcystin blooms is available. Analysis of these maps can lead to the discovery of yearly problem spots, factors that contribute to microcystin generation, and probable directions of travel for the blooms.

IV. Results and Discussion

A. Method Reliability

An IDFM-based early warning system for quantifying toxin levels in algal blooms using satellite remote sensing data depends upon two primary constituents for success: 1) accurate surface reflectance data of the water body; and 2) a reliable algorithm for predicting microcystin concentration. This section will quantify the advantages of using enhanced spatial data (MERIS and MODIS ocean color bands) over less detailed products (Landsat and MODIS land bands) using traditional two band inversion models and more computationally-intensive GP models. As detailed in Table II, 44 ground-truth samples were used to train and calibrate the models. 60% of the input data was used to train the models, and the remaining 40% was used to validate the performance of the model. The method for splitting up the data into training and calibration sets is as follows: 1) order the ground-truth values from low to high; and 2) in an alternating manner, assign data to the training, and validation data sets. This procedure exposes the models to the same range of microcystin concentration values during training and validation.

While the traditional two-band models will always yield the same coefficients when solved using regression techniques, the equation and performance of each GP model will vary during different runs. This is a result of the random starting weights and the fundamental methodology used during model creation. To lucidly depict the variation and average performance of the GP models, five separate training runs were performed using Discipulus for both multispectral and hyperspectral inputs. Only the single best model from each training run was selected for analysis. The second point of interest was to determine whether the models take seconds, minutes, or hours to train on average. Computational time is also dependent upon the computer hardware used to run Discipulus. During each training run, Discipulus was the only program actively open and used on the computer. The computer specifications are as follows: Intel Core i7-3720QM CPU at 2.6 GHz, 24 574 MB RAM, and 500 GB hard drive. Finally, a comparison is made between data sets derived from the multispectral and hyperspectral sensors. The coefficient of determination, time required to computing each model, and the run number of each model are detailed in Table III.

The GP models using fused hyperspectral sensor data products took 322 s longer to solve on average, yet the resulting coefficient of determination was 0.8883, which is 0.0717 greater than the coefficient of determination derived from fused multispectral sensor data products. The multispectral sensor GP model yielded a less accurate fitness of 0.1652 compared to the hyperspectral sensor GP model's fitness of 0.1425. The multispectral solutions had shorter run times, since they stopped improving earlier on in the model development. The greater coefficients of determination for the hyperspectral sensor GP models is attributed to the finer band widths (refer to Fig. 1), which allows for telltale peaks and troughs of chlorophyll-a and phycocyanin (indicators of microcystin) to be more readily identified. An interesting observation is made by analyzing the run times for the multispectral sensor GP models. The fifth model was derived in a mere 5 s, while the next closest solution was obtained in 92 s. The order of magnitude difference is rare, yet it is due to randomly generated starting weights and randomly selected input data that are used to initiate model formulation. In this case, the program stumbled upon an excellent combination of parameters for determining the relationship between surface reflectance and microcystin concentration.

Verifying that the model has appropriately related the band data from the fused images to the microcystin ground-truth data is accomplished through a least squares analysis between the observed and the predicted microcystin values. The best model is selected based on the coefficient of determination, fitness level

TABLE III
STATISTICAL COMPARISON BETWEEN GP MODELS CREATED USING FUSED DATA FROM MULTISPECTRAL AND HYPERSPECTRAL SENSORS

	Model number	1	2	3	4	5	**AVG**
Fused multispectral sensor GP models	R^2	0.8425	0.7931	0.7683	0.8449	0.8344	**0.8166**
	Fitness	0.1423	0.1605	0.2256	0.1326	0.1648	**0.1652**
	Run time (s)	194	272	246	92	5	**162**
	Run number	34	43	40	26	4	**29**
Fused hyperspectral sensor GP models	R^2	0.8243	0.9269	0.8847	0.9177	0.8879	**0.8883**
	Fitness	0.1629	0.1454	0.1276	0.1764	0.1002	**0.1425**
	Run time (s)	211	450	437	932	389	**484**
	Run number	30	50	48	71	43	**48**

achieved, and a visual confirmation that the model can accurately identify peak microcystin values. Identification of high microcystin values is imperative for an early warning system. Based on these criteria, the fourth fused multispectral sensor GP model and the second fused hyperspectral sensor GP model from Table III were selected for further analysis. The predictive capabilities of 3 GP models developed from pure MERIS [(a) and (d)], fused multispectral sensor data [(b) and (e)], and fused hyperspectral sensor data [(c) and (f)] are presented in Fig. 7.

The MERIS GP model [Fig. 7(a) and (d)] served as a reference to compare the fused GP models (the two GP models derived using the fused spectral data as inputs) to. Only 26 data points were available to generate the MERIS model, compared to 41 data points for the fused models; yet the pure MERIS model actually had the best performance with a coefficient of determination equal to 0.9469 and admirable capacity at predicting peak values. More data points were available for the fused models, due to the additional information provided by fusion with MODIS. The MERIS GP model obviously requires less computational power and data storage requirements to develop, since data from only one satellite sensor is necessary.

In analyzing the multispectral [Fig. 7(b) and (e)] and hyperspectral [Fig. 7(c) and (f)] performance, the GP model created from the hyperspectral sensor product yields a better coefficient of determination of 0.9269 compared to 0.8449. Both of the models are capable of predicting peak microcystin values, which is a necessary function for delineating between a HAB laced with microcystin and an algal bloom comprised of nontoxic algal species. The predictive capabilities of the fused hyperspectral sensor GP model excelled at predicting microcystin concentrations less than 1 μgL^{-1}, while the multispectral model simply plateaus in this region as detailed in Fig. 7(b). The difference in predictive power at low concentrations is also identified when comparing Fig. 7(e) and (f). As seen by the horizontal set of data points in Fig. 7(e), the fused multispectral sensor GP model consistently underestimates low microcystin values. The fused hyperspectral sensor GP model shown by Fig. 7(f) has the advantage of more accurately predicting low microcystin values, which allows for the identification of HAB formation at an early stage. As a result, these areas can be more closely monitored for continued HAB formation. This is also useful for assessing water quality in environmentally sensitive areas. The near real-time early warning system with more accurate microcystin prediction at all concentrations is paramount, and the GP model formulated with input from the hyperspectral sensor successfully achieved this.

B. Model Predictability

1) Performance Comparison Between Traditional Inversion Modeling and Data Mining: To compare predictability between the GP models and traditional inversion methods, a two band ratio model and a spectral slope model were used [13], [37]. The ideal bands for the two band models were found by testing all possible band combinations and choosing the two bands that yielded the highest coefficient of correlation and fitness. For the GP models, all of the bands were supplied as inputs to Discipulus, and the program determined the bands that shared a relationship with the microcystin concentration. Comparing the two band models along with the GP models was done using four statistical indices: the root-mean-square error (RMSE), ratio of standard deviations (CO), mean percent error (PE), and the square of the Pearson product moment correlation coefficient ($\mathrm{RSQ} = \mathrm{R}^2$). The results are presented in Table IV and with special attention to the computational time required to solve the models:

The observed microcystin mean values are the same for each of the models, since they share the same set of ground-truth data. The next point of detail is that the traditional two-band models performed worse than the spectral slope and GP models. The spectral slope models performed poorly when using multispectral input values ($\mathrm{R}^2 = 0.09625$), but this model performed significantly better for the hyperspectral sensor surface reflectance inputs ($\mathrm{R}^2 = 0.7062$). This is likely due to the enhanced spectral detail. As previously mentioned, the multispectral sensor data covers a wide portion of the electromagnetic spectrum for each band. This often leads to spectral peaks and troughs becoming averaged with nearby data; thus, losing their shape and detail. The fused hyperspectral sensor product is better suited for the spectral slope model for microcystin prediction, since the defining features in the spectral reflectance curves for chlorophyll-a and phycocyanin are captured. As a result, the clearly delineated peak or trough values produce a lucid response when analyzed with the spectral slope model. The next comparison is focused on the GP models. The hyperspectral sensor GP model underestimates the mean observed microcystin value by

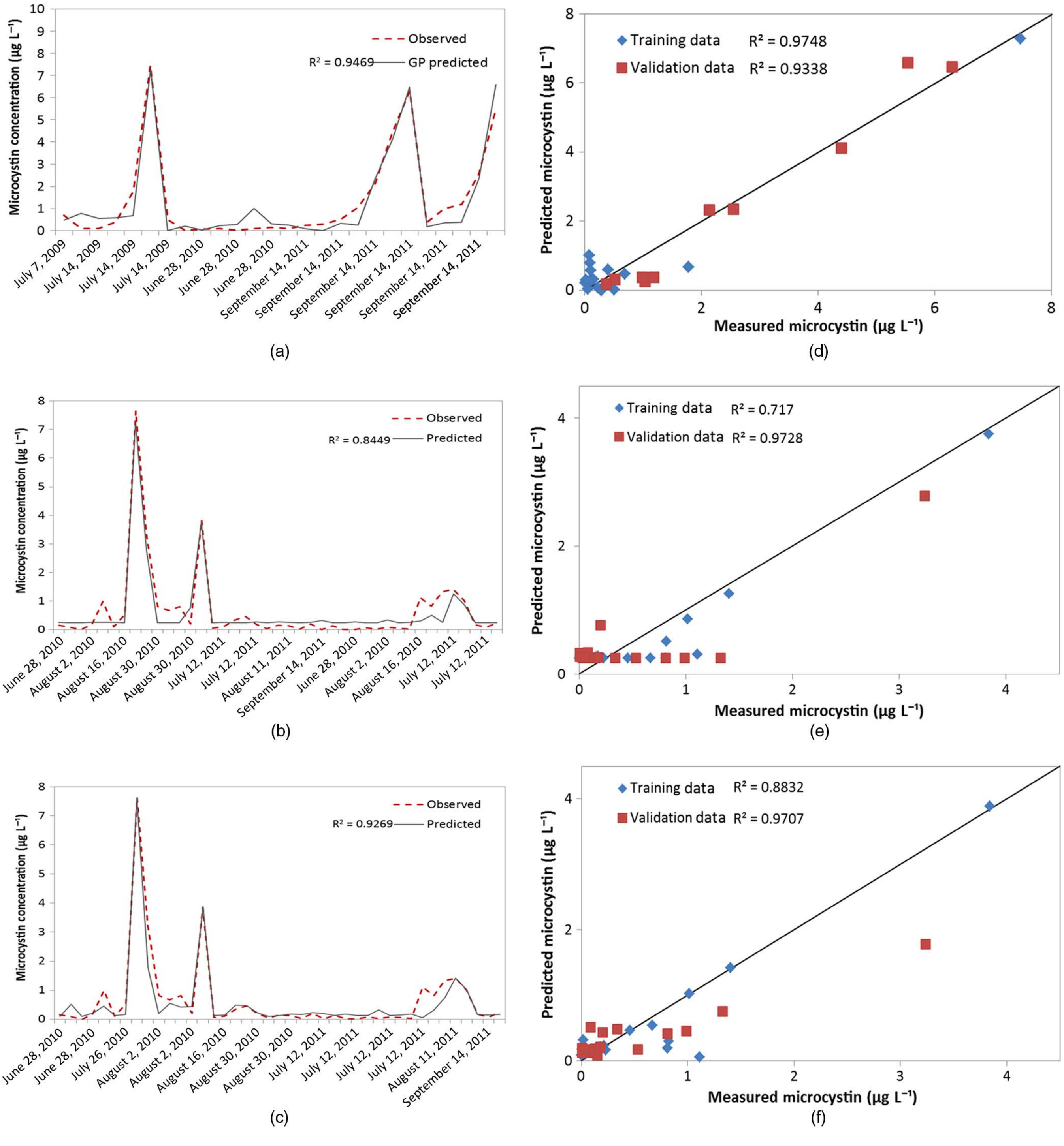

Fig. 7. Time-series graphs in the left column exhibit the predictive performance of a pure MERIS GP model (a), a fused multispectral sensor GP model (Landsat and MODIS land bands) (b), and a fused hyperspectral sensor GP model (MERIS and MODIS ocean color bands) (c). As can be seen in images (a)–(c), the models aptly predict peak microcystin values; however, the ability to predict low-microcystin values varies between the models. In (a), it can be seen that the predicted values at low concentrations show mediocre correlation with the observed values. From image (b), the model has a horizontal line for predicting observed values below $0.3\ \mu\mathrm{gL}^{-1}$. The hyperspectral sensor GP model (c) having the best success at estimating the microcystin at low concentrations. For images in the right column (d)–(f), the predicted microcystin values have been plotted against the observed values to accentuate any biases that are present in the predicted values.

$0.0782\ \mu\mathrm{gL}^{-1}$, while the multispectral sensor GP model is much closer to the mean with an average underestimation of $0.0338\ \mu\mathrm{gL}^{-1}$. Recall from Fig. 7(b) that the multispectral model often overpredicted the microcystin values at low concentrations, yet it underpredicts values close to $1\ \mu\mathrm{gL}^{-1}$. The hyperspectral sensor model matched the observed microcystin values more closely, but this model underestimated concentrations more often than it overestimated concentrations. The GP model using the hyperspectral sensor yielded an RMSE value of $0.3530\ \mu\mathrm{gL}^{-1}$, which is slightly worse than the $0.3451\ \mu\mathrm{gL}^{-1}$ obtained by the multispectral sensor GP model. Both of these minor shortcomings for the hyperspectral sensor GP model are compensated for by the improved performance with regard to the CO, PE, and R^2 values. The multispectral and hyperspectral GP

TABLE IV
GP and Two-Band Models Using Multispectral and Hyperspectral Sensor Surface Reflectance Input Data Are Evaluated Using Four Indices of Accuracy. Bolded Values Indicate the Two Models Exhibiting the Highest Performance in the Assigned Statistical Category

	Ideal value	Fused multispectral sensor input*			Fused hyperspectral sensor input**		
		Two-band ratio model	Spectral slope model	GP model	Two-band ratio model	Spectral slope model	GP model
Observed microcystin mean (μg L^{-1})	–	0.6718	0.6718	0.6718	0.6718	0.6718	0.6718
Predicted microcystin mean (μg L^{-1})	–	2.226	0.1792	0.6360	1.008	0.3571	0.5936
Root-mean-square error (μg L^{-1})	0	1.348	1.340	**0.3451**	1.356	0.7583	**0.3530**
Ratio of St. Dev.	1	0.8270	0.1238	**0.6787**	0.5540	0.5589	**0.6837**
Mean percent error (%)	0	87.57	**5.251**	38.07	61.87	**2.177**	25.01
Square of the Pearson product Moment correlation coefficient	1	0.02393	0.09625	**0.8449**	0.2710	0.7062	**0.9269**
Computational time (s)	–	< 1	< 1	92	< 1	< 1	450

The computational time is the amount of seconds required to generate the model. As expected, machine learning methods took longer to solve than regression techniques. The fused hyperspectral input provided the most accurate results overall ($N = 44$).

*Fused multispectral sensor input pair: Landsat and MODIS land bands.
**Fused hyperspectral sensor input pair: MERIS and MODIS ocean color bands.

TABLE V
Comparison Between GP Model Performance From Single Sensor Inputs and Fused Sensor Inputs

	Single sensor		Fused sensor pair	
	MODIS GP model	MERIS GP model	Multispectral GP model	Hyperspectral GP model
Root-mean-square error (μg·L^{-1})	0.1351	0.4951	0.3451	0.3530
Ratio of standard deviations	0.6926	0.7127	0.6787	0.6837
Mean percent error (%)	23.80	5.809	38.07	25.01
Square of the Pearson product moment correlation coefficient	0.9204	0.9469	0.8449	0.9269
Spatial resolution (m)	1000	300	30	300
Temporal resolution (days)	1	3	1	1

On average, the GP models created using fused sensor pairs exhibited slightly lower accuracy, yet the temporal and spatial resolution of the fused sensor GP model's outputs are superior to single sensor GP model's outputs.

models had CO values of 0.6787 and 0.6837, which are close to the ideal value of 1. The hyperspectral GP model yielded a PE of 25%, while the multispectral sensor GP model had 13.06% higher error. With regards to the R^2 values, both models exhibited strong statistical significance and a positive correlation with values of 0.8449 for the multispectral sensor GP sensor model and 0.9269 for the hyperspectral sensor GP model. In conclusion, the GP models are better suited for determining the complex, nonlinear relationship between microcystin and surface reflectance, and hyperspectral surface reflectance inputs yielded more accurate results than multispectral sensor surface reflectance inputs.

Analysis of the computational time required to derive the models provides interesting theoretical insights. The drawback for using machine learning techniques is the time required to retrieve and formulate the nonlinear models. For the purpose of comparison, the traditional models were also solved using regression techniques. Computational time required for solving the multispectral and hyperspectral sensor GP models are 92 and 450 s, respectively. The hyperspectral spectral slope model actually yielded reasonable predictions for microcystin. The model trained in less than a second, compared to the 92 and 450 s training times for the GP models. The short training time for the hyperspectral spectral slope model is a minor benefit, yet in the context of this study in which the GP models took 29 and 484 s to train on average (Table III), the training time advantage is negligible. Furthermore, the task of training an accurate GP model only needs to be performed once, with periodic updates in the future, for determining the relationship between a water body's unique surface reflectance characteristics and the microcystin levels.

2) Impact of Data Fusion on GP Model Results: During data fusion, the spectral integrity of the data is altered. To determine whether the data from the fused sensor GP models are reliable, a statistical comparison is offered against GP models derived from single sensor data, as is shown in Table V.

The GP models using single sensor data each had similar or superior accuracy than the GP models trained from fused sensor data. But, before the fused GP models are characterized as underperforming, their inherent advantages should be discussed. The revisit time of MERIS is up to 3 days in length, which leaves sizeable data gaps. The MODIS GP model does not suffer from the longer revisit time of MERIS, yet it has a resolution of 1000 m compared to 300 m of MERIS, which provides poorer insight into HAB identification and delineation. Even though the fused GP models offer slightly lower performance, their prediction capabilities are still quite strong, as is shown by the statistical analysis in Table V. Furthermore, the fused GP models offer a superior early warning system, because they are able to provide medium to high-resolution data for generating microcystin concentration maps on a daily basis. In conclusion, the drawback of additional computing power and storage capacity required

TABLE VI
SPECTRAL BAND CENTERS WITH THE HIGHEST PERFORMANCE FOR THE TRADITIONAL TWO-BAND MODELS

Model type	Band centers (nm)	R^2
Multispectral two-band ratio	570 and 840	0.02393
Multispectral spectral slope	570 and 660	0.09625
Hyperspectral two-band ratio	560 and 681	0.2710
Hyperspectral spectral slope	665 and 681	0.7062

TABLE VII
FREQUENCY OF USE FOR THE BAND CENTERS USED AS SPECTRAL INPUTS FOR THE MULTISPECTRAL AND HYPERSPECTRAL SENSOR GP MODELS

Fused multispectral sensor input		Fused hyperspectral sensor input	
Band center (nm)	Frequency of use (%)	Band center (nm)	Frequency of use (%)
485	67	**412**	**83**
560	53	**443**	**80**
660	**97**	490	57
840	57	560	27
1650	**80**	665	7
2090	**70**	**681**	**100**

The top 3 bands for each sensor type have been bolded.

to formulate the fused hyperspectral sensor GP models is outweighed by the benefits of daily, 300 m concentration maps with comparable performance.

3) Analysis of Spectral Importance in Model Development: Additional insights can be gleaned by analyzing the spectral bands that were used to create each of the models. The bands used to train the two-band models are provided in Table VI.

The band centers most used by the traditional two-band models fall in the range of 560–570 nm and 660–681 nm. This corresponds to the spectral features observed in Fig. 1. Chlorophyll produces a distinct reflectance dip or trough in the range 660–681 nm, and it is clear why this wavelength would exhibit a strong relationship with microcystin prediction. Next, the GP models were analyzed for the frequency of use for each of the six bands (Table VII). The frequency of use is how often a specific band was used in the 30 best programs or models created by Discipulus. Hundreds of millions of programs were generated by Discipulus while solving for the 30 best models; yet the majority of the programs developed exhibit unsatisfactory performance. As a result, only the 30 best models are used to infer scientific observations because of their accuracy. If a band was used in every program, it would have 100% frequency of use, and it would likely share a high correlation between surface reflectance and microcystin. The frequency of use for the variables identified in the GP models is presented in Table VII.

The fused multispectral and hyperspectral sensor inputs do not share many of the same band centers, so a direct comparison cannot be made between the two sensor types, since they observed different portions of the electromagnetic spectrum. Nevertheless, an independent analysis of the frequency of use can be offered for each sensor type. The fused multispectral sensor GP model favored band centers at 660, 1650, and 2090 nm. This corresponds to bands 3, 5, and 7 of Landsat and the MODIS land color bands 1, 6, and 7. Comparison to Fig. 1 shows that the wide band center at 660 nm averages spectral features unique to phycocyanin and chlorophyll-a (both indicators of microcystin). The strong emphasis placed on the shortwave infrared bands for predicting microcystin can be attributed to the decrease in absorption at these bands as the chlorophyll content in the water increases. The fused hyperspectral sensor GP model frequency used bands centered at 412, 443, and 681 nm; 412 nm plays an important role for detecting decaying matter from once living organisms. This is directly tied to observed microcystin concentrations in the water, because *Microcystis* produce the toxin microcystin at the end of their exponential growth phase, and the toxin is retained within the bacteria until death. As the bloom dies off, the cell walls rupture and release the toxin into the water. The band center at 443 nm is the chlorophyll absorbance maximum. The band center at 681 nm was also used in the traditional two-band models, as it directly corresponds with a strong reflectance trough caused by chlorophyll-a in the water. While phycocyanin and chlorophyll-a serve as strong microcystin indicators, variations in optical complexity, such as heavy suspended solid levels commonly induced by storm events in the Maumee Bay region, may require the identification of more abstract indicators, such as HAB growth rate (tied to microcystin production [9]) and weather patterns, which may limit light levels.

C. Microcystin Maps

Using the GP model derived from the fused band data, maps of the microcystin concentration throughout Lake Erie can be reconstructed to allow for the assessment of blooms during the summer. As a result, detailed information on *Microcystis* bloom proliferation and transportation can be identified and subsequently used to identify probable problem spots that require close monitoring during the summer. To illustrate this, microcystin maps generated from a fused hyperspectral sensor GP model and a fused multispectral sensor GP model are shown in Fig. 8, and they are compared to a false color image of the algal bloom occurring on the same day:

The concentration map derived using the fused data with hyperspectral sensor inputs [Fig. 8(a)] is much less detailed, due to the 300 m resolution. The apparent advantage is that the predicted medium and high concentrations of microcystin align with the green HABs observed in the true color image [Fig. 8(c)]. The less accurate concentration map derived from multispectral sensor data [Fig. 8(b)] appears to exaggerate microcystin concentrations throughout the lake. The high likelihood for false positives to occur greatly diminishes the usefulness of the early warning system based on the fused multispectral sensor data. Given this performance, it would appear that the bands from the hyperspectral sensor fusion pair are more useful for differentiating between open water and HABs. However, the enhanced detail provided by the 30 m resolution yields insight into the benefits that a hyperspectral satellite with fine spatial resolution would yield. This is evidenced by the apparent currents and potential bloom delineations seen in Fig. 8(b). Enhancing the spatial resolution of the MERIS sensor could theoretically be performed by fusing the MERIS images with another sensor that has a higher spatial resolution. Should this

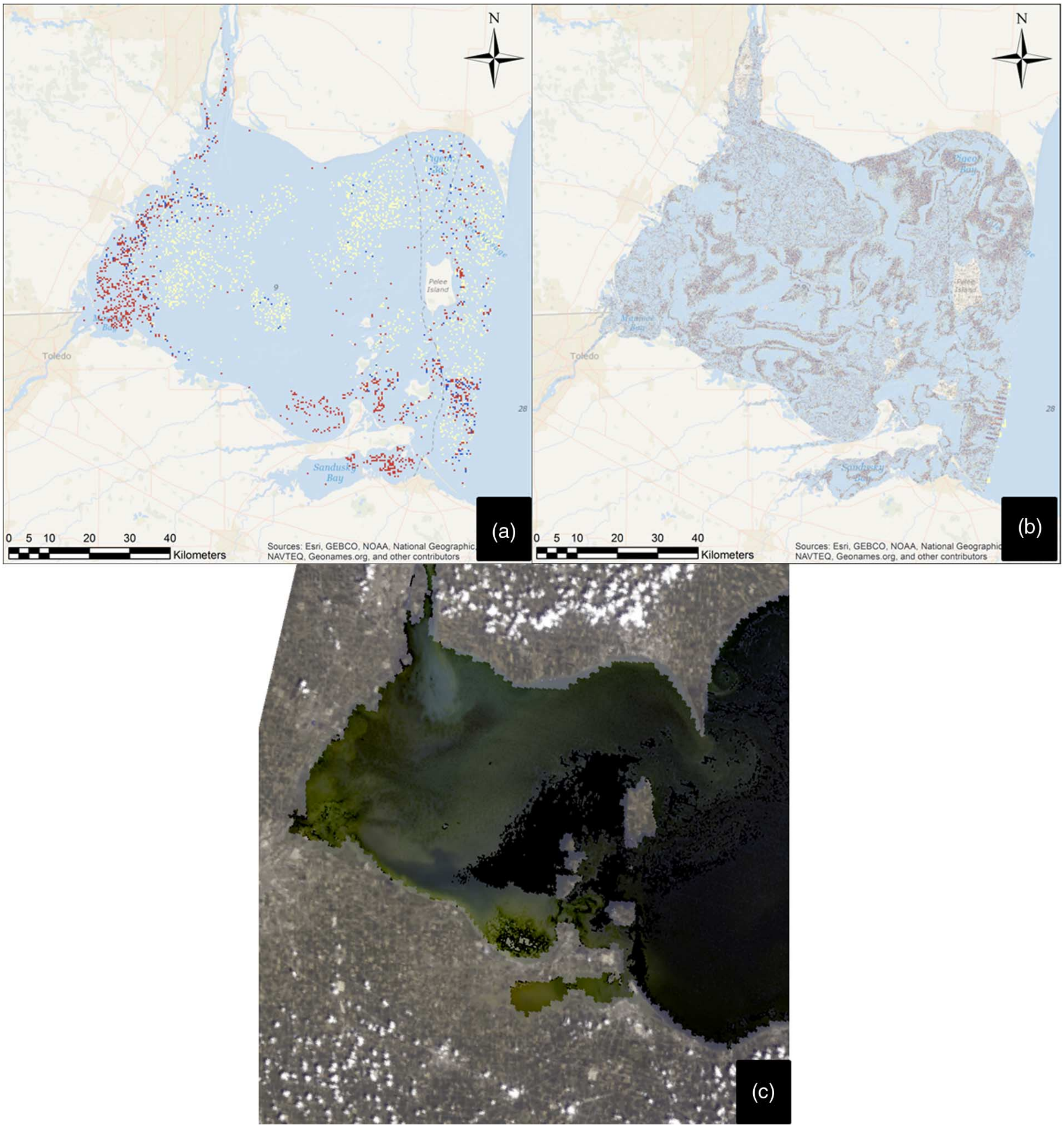
N
Toledo
Pelee Island
Sandusky Bay
0 5 10 20 30 40 Kilometers
Sources: Esri, GEBCO, NOAA, National Geographic, NAVTEQ, Geonames.org, and other contributors
(a)
N
Toledo
0 5 10 20 30 40 Kilometers
Sources: Esri, GEBCO, NOAA, National Geographic, NAVTEQ, Geonames.org, and other contributors
(b)
(c)

Fig. 8. Concentration maps were generated using the hyperspectral sensor GP model (a) and multispectral GP model (b). The false color image of western Lake Erie is presented on the bottom (c). Large algal blooms spawning out of the Maumee and Sandusky Bays on July 26, 2010 are seen as dark green, while the sediment is a pale white in (c). Dark red spots in (a) and (b) denote areas of high microcystin concentration that pose a health threat, while yellow spots indicate low to medium concentrations. The 30 m spatial resolution of the multispectral image provides more detailed outlines, while the coarser (300 m) hyperspectral resolution predicts microcystin concentrations in locations that more closely align with HAB presence.

task be carried out, there are a number of concerns that must be addressed. The satellite pair must share common spectral bands, as was previously detailed in Table I. Additionally, if MERIS is fused with a satellite of higher spatial resolution and the resulting image is then fused with MODIS to improve the temporal resolution, then the spectral errors and fusion approximations made during the first fusion step will propagate when fusing the second time to improve the temporal resolution. Finally, the mission of the Envisat satellite housing the MERIS sensor ended in April 2012 after a loss of contact; yet a suitable replacement is Sentinel-3 mission slated for launch in 2015 by the European Space Agency. Sentinel-3 similarly shares a 300 m spatial resolution, yet its spectral capabilities are higher at 21 bands compared to 15 of MERIS.

V. Conclusion

STARFM was able to accurately fuse the both hyperspectral (MERIS and MODIS ocean color bands) and multispectral (Landsat and MODIS land bands) image pairs to generate synthetic images possessing both moderate spatial and temporal resolution. The synthetic images contain more data than a single image from either satellite, and the fusion method is used to fill in data gaps from the lengthy revisit times of MERIS and Landsat. In comparing traditional two-band models to more complex GP models, it was observed that the GP models required longer training times, yet they offered higher explanatory power in relating microcystin to surface reflectance. Next, it was shown that the fused hyperspectral sensor GP model excelled over the fused multispectral sensor GP model for microcystin prediction. This was quantified using four statistical indices. The fused multispectral sensor GP model yielded more desirable mean prediction errors and RMSE values of 0.0358 and $0.3451\ \mu\mathrm{gL}^{-1}$ compared to 0.0782 and $0.3530\ \mu\mathrm{gL}^{-1}$. The fused hyperspectral sensor GP model ranked the highest when evaluated with the CO, PE, and R^2 statistical indices, achieving values of 0.6837, 25.01%, and 0.9269, compared to 0.6787, 38.07%, and 0.8449, respectively. While the fused hyperspectral sensor GP model required the longest training time of 7.5 min, it had the highest explanatory power microcystin prediction and the fulfillment of an early warning system.

One limiting factor to the ground-truth data is that the majority of the samples correspond to fixed points that were sampled when HABs on the lake were observed. Ideally, sampling would have been carried out on a daily basis starting from when the HAB formed and stopping after it dissipated. This would provide a representative idea on when microcystin began to form within the HAB, and daily tracking of the HAB with corresponding microcystin samples could corroborate the success of such a map, since a time series of maps would lucidly depict HAB mobility. The second limitation is that many of the ground-truth points were below $1\ \mu\mathrm{gL}^{-1}$. Even though the GP models successfully predicted peak microcystin concentrations, a larger and more diverse data set would improve the predictability of the models at low and peak values. With the recent failure of the MERIS sensor, this work would be further explored following the upcoming launches of the Sentinel multi-satellite project.

References

[1] World Health Organization (WHO), *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management*. London, U.K.: E & FN Spon, 1999.

[2] J. Lekki, R. Anderson, Q. Nguyen, and J. Demers, "Development of hyperspectral remote sensing capability for early detection and monitoring of harmful algal blooms (HABs) in the Great Lakes," in *Proc. AIAA Aerospace Conf.*, Seattle, WA, USA, Apr. 6–9, 2009.

[3] B. Hitzfeld, S. Hoger, and D. Dietrich, "Cyanobacterial toxins: Removal during drinking water treatment and human risk assessment," *Environ. Health Perspect.*, vol. 108, no. 1, pp. 113–122, Mar. 2000.

[4] D. Toivola, J. Eriksson, and D. Brautigan, "Identification of protein phosphate 2A as the primary target for microcystin-LR in rat liver homogenates," *FEBS Lett.*, vol. 344, no. 2–3, pp. 175–180, May 1994.

[5] G. Jones and P. Orr, "Release and degradation of microcystin following algaecide treatment of a Microcystis aeruginosa bloom in a recreational lake, as 248 determined by HPLC and protein phosphate inhibition assay," *Water Resour.*. vol. 28, pp. 871–876, 1994.

[6] M. Rogalus and M. Watzin, "Evaluation of sampling and screening techniques for tiered monitoring of toxic cyanobacteria in Lakes," *Harmful Algae*, vol. 7, no. 4, pp. 504–514, Jun. 2008.

[7] J. Rinta-Kanto *et al.*, "Lake Erie Microcystis: Relationship between microcystin production, dynamics of genotypes and environmental parameters in a large lake," *Harmful Algae*, vol. 8, no. 5, pp. 665–673, Jun. 2009.

[8] J. Budd, A. Beeton, R. Stumpf, D. Culver, and W. Kerfoot, "Satellite observations of Microcystis blooms in western Lake Erie," *Verh. Int. Verein. Limnol.*, vol. 27, pp. 3787–3793, 2001.

[9] T. Wynne *et al.*, "Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes," *Int. J. Remote Sens.*, vol. 29, no. 12, pp. 3665–3672, 2008.

[10] G. Ganf, R. Oliver, and A. Walsby, "Optical properties of gas-vacuolate cells and colonies of Microcystis in relation to light attenuation in a turbid, stratified reservoir (Mount Bold Reservoir, South Australia)," *Mar. Freshwater Res.*, vol. 40, no. 6, pp. 595–611, 1989.

[11] J. Mole, C. Chow, M. Drikas, and M. Burch, "The influence of cultural media on growth and toxin production of the cyanobacterium Microcystis aeruginosa Kutz Emend Elenkin," presented at the *13th Annu. Conf. Aust. Soc. Psychol. Aquat. Bot.*, Hobart, Australia, Jan. 1997.

[12] J. Ha, T. Hidaki, and H. Tsuno, "Analysis of factors affecting the ratio of microcystin to chlorophyll-a in cyanobacterial blooms using real-time polymerase chain reaction," *Environ. Toxicol.*, pp. 21–28, 2009, doi: 10.1002

[13] R. Vincent *et al.*, "Phycocyanin detection from Landsat TM data for mapping cyanobacterial blooms in Lake Erie," *Remote Sens. Environ.*, vol. 89, no. 3, pp. 381–392, Feb. 2004.

[14] W. Yuchun and C. Wei, "Landsat TM image feature extraction and analysis of algal bloom in Taihu Lake," in *Proc. SPIE*, 2008, vol. 7000.

[15] European Space Agency (ESA), *MERIS Product Handbook Issue 2.1*, Paris, France: ESA, 2006.

[16] P. Pabich, *Hyperspectral Imagery: Warfighting Through a Different Set of Eyes*. AL, USA: Maxwell Air Force Base, Air University, 2002.

[17] P. Shippert, *Introduction to Hyperspectral Image Analysis*. Norwalk, CT, USA: Research Systems, Inc., n.d.

[18] A. Del Bianco, G. Serafino, and G. Spock, *An Introduction to Spectral Imaging*. Magdalen, Austria: Carinthian Tech Research GmbH, n.d.

[19] W. Belokon *et al.*, *Multispectral Imagery Reference Guide*. Fairfax, VA, USA: Gogicon Geodynamics, 1997.

[20] G. Chang *et al.*, "The new age of hyperspectral oceanography," *Oceanography*, 2004.

[21] J. O'Reilly *et al.*, "Ocean color chlorophyll algorithms for SeaWiFS," *J. Geophys. Res.*, vol. 103, no. 24, pp. 937–953, 1998.

[22] C. Hu, K. Carder, and F. Muller-Karger, "Atmospheric correction of SeaWiFS imagery over turbid coastal waters: A practical method," *Remote Sens. Environ.*, vol. 74, no. 2, pp. 195–206, Nov. 2000.

[23] Z. Lee and K. Carder, "Effect of spectral band numbers on the retrieval of water column and bottom properties from ocean color data," *Appl. Opt.*, vol. 41, no. 12, pp. 2191–2201, Apr. 2002.

[24] T. Bridgeman, "*Water quality monitoring in Western Lake Erie and Maumee Bay*," Univ. Toledo Lake Erie Center, Lake Erie Protection Fund (Project LEPF 03-19), Oregon, OH, USA, 2005.

[25] B. Lubac *et al.*, "Hyperspectral and multispectral ocean color inversions to detect Phaeocystis globosa blooms in coastal waters," *J. Geophys. Res.*, vol. 113, no. C6, Jun. 2008.

[26] E. Torrecilla, J. Piera, and M. Vilaseca, *Derivative Analysis of Hyperspectral Oceanographic Data, Advances of Geoscience and Remote Sensing*, G. Jedlovec, Ed. Vienna, Austria: InTech, 2009, ISBN: 978-953-307-005-6.

[27] A. Ouellette, S. Handy, and S. Wilhelm, "Toxic microcystis is widespread in Lake Erie: PCR detection of toxin genes and molecular characterization of associated cyanobacterial communities," *Microb. Ecol.*, vol. 51, no. 2, pp. 154–165, Feb. 2006.

[28] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006.

[29] A. Michalak *et al.*, "Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions," in *Proc. Nat. Acad. Sci.*, 2013, vol. 110, pp. 6448–6452.

[30] E. Vermote, S. Kotchenova, and J. Ray, *MODIS Surface Reflectance User's Guide. Version 1.3.* College Park, MD, USA: MODIS Landsat Surface Reflectance Science Computing Facility, 2011.

[31] United States Geological Survey (USGS). (2014). *Landsat Processing Details* [Online]. Available: http://landsat.usgs.gov/Landsat_Processing_Details.php

[32] E. Vermote, N. Saleous, and C. Justice, "Atmospheric correction of MODIS data in the visible to middle infrared: First results," *Remote Sens. Environ.*, vol. 83, no. 1/2, pp. 97–111, Nov. 2002.

[33] J. Masek *et al.*, "A Landsat surface reflectance data set for North America, 1990–2000," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 68–72, Jan. 2006.

[34] J. Van Genderen and C. Pohl, "Image fusion: Issues, techniques, and applications. Intelligent image fusion," in *Proc. EARSel Workshop*, Strasbourg, France, Sep. 1994, pp. 18–26.

[35] C. Pohl and J. Van Genderen, "Multisensor image fusion in remote sensing: Concepts, methods, and applications," *Int. J. Remote Sens.*, vol. 19, no. 5, pp. 823–854, 1998.

[36] D. Francone, *Discipulus Software Owner's Manual, Version 3.0 DRAFT*, CO, USA: Machine Learning Technologies, Inc., 1998.

[37] P. Dash *et al.*, "Estimation of cyanobacterial pigments in a freshwater lake using OCM satellite data," *Remote Sens. Environ.*, vol. 115, no. 12, pp. 3409–3423, Dec. 2011.

Ni-Bin Chang received the B.Sc. degree in civil engineering from National Chiao-Tung University, Hsinchu City, Taiwan, and the M.Sc. and Ph.D. degrees in environmental systems engineering from the Cornell University, Ithaca, NY, USA.

He is a Professor with the University of Central Florida, Orlando, FL, USA.

Benjamin Vannah received the B.Sc. degree in mechanical engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2009, and the M.Sc. degree in environmental engineering from the University of Central Florida, Orlando, FL, USA, in 2013.

Currently, he is a Graduate Research Assistant with the University of Central Florida.

Y. Jeffrey Yang received the B.Sc. degree in geomechanics from China University of Geosciences, Beijing, China, and the M.Sc. degree in geochemistry from Chinese Academy of Geological Sciences, Beijing, China, and the Ph.D. degree in isotope geochemistry from Miami University, Oxford, OH, USA.

He is a Research Scientist with the Office of Research and Development, EPA, USA.